

# Architektura dużych projektów bioinformatycznych

Bartek Wilczyński

[bartek@mimuw.edu.pl](mailto:bartek@mimuw.edu.pl)

<http://www.mimuw.edu.pl/~bartek>

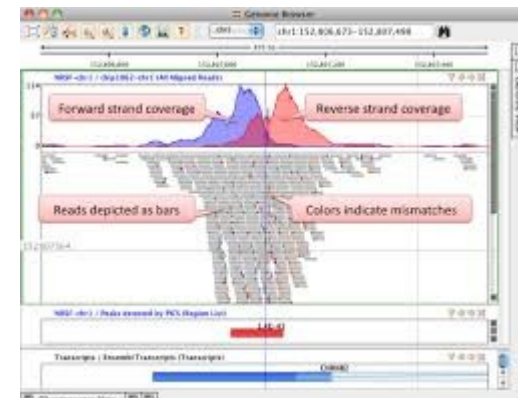
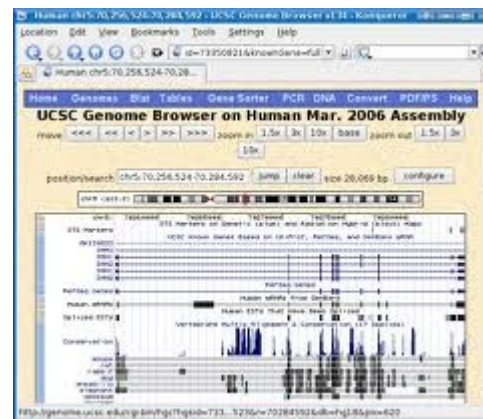
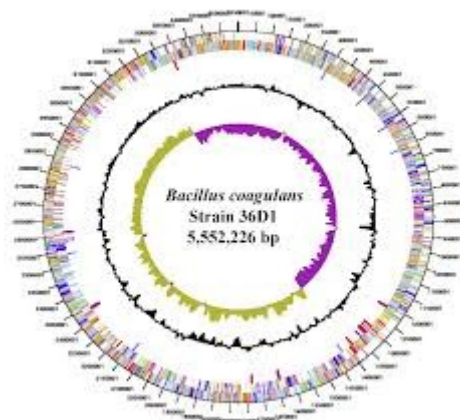
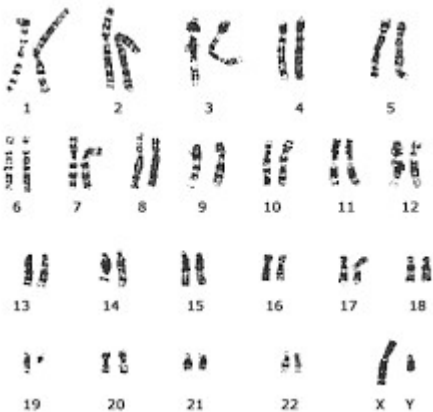
**Wykład 7. - Przeglądarki genomów**  
15.IV. 2020

# Tematy na dziś

- Po co są przeglądarki genomowe
- Co zawierają opisy genomów (annotation)
- UCSC genome browser
- Gbrowse
- ENSEMBL
- IGB, IGV i podobne
- GenomeDiagram i podobne

# Przeglądarki genomowe

- Od czasu, kiedy dostępne są całe genomy organizmów zachodzi potrzeba wizualizacji
- Genomy bakteryjne często były wizualizowane w całości
- Wraz z pojawieniem się dużych genomów zaszła potrzeba innej wizualizacji
- Teraz także duże zbiory danych stawiają nowe wymagania

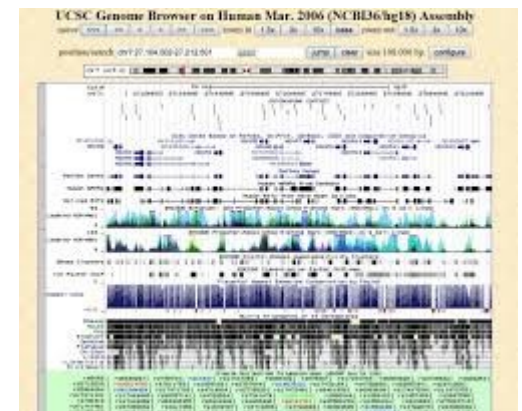
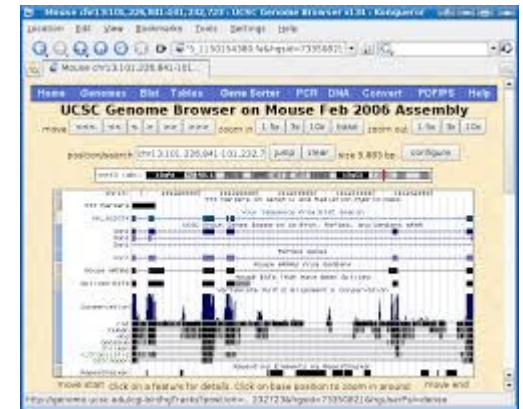


# Historycznie

- 1976 - Pierwszy genom RNA bakteriofag MS2
- 1977 – Pierwszy genom DNA (5386 bp)
- 1995
  - *Haemophilus Influenzae* – bakteria 1.8M bp
  - *Saccharomyces Cerevisiae* – eukariont – 12.1 M bp
- 1996 – archea *Methanocaldococcus jannaschii*
- 1997 – *E. coli*
- 2000 – *H. sapiens*
- 2020 – >590tys sekwencji typu “whole genome” w genbank (>240tys bakteryjnych, >112tys wirusów, >97tys zwierzęcych, >46 tys. Roślinnych, etc.)

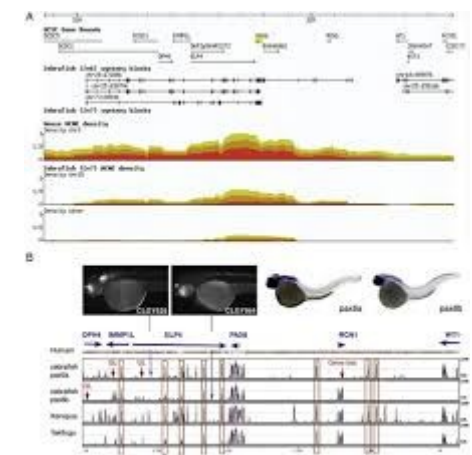
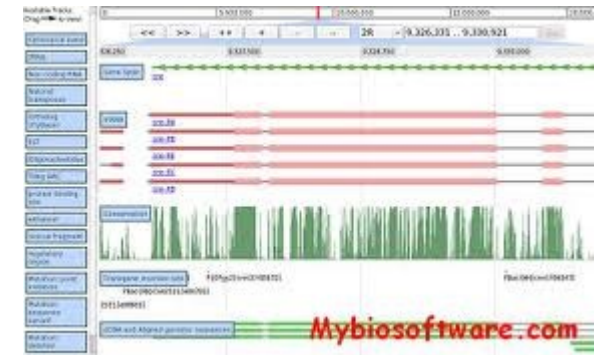
# UCSC browser

- Jim Kent, 2000, napisany w C na potrzeby publikacji genomu ludzkiego
- Udostępniany darmowo dla akademickich zastosowań
- Komercyjna licencja Kent Informatics
- Ciągłe jedna z bardziej użytecznych przeglądarek
- W zasadzie brak możliwości dostosowania do własnych potrzeb, dobre dla użytkowników, ale nie dla developerów



# Gbrowse

- Część projektu GMOD
- Rozpoczęty w 2002
- PERL artistic license
- Rozwijany głównie w PERLu
- Coraz więcej java-scriptu, począwszy od 2007r.
- Ogromna liczba instalacji
- Jbrowse, WebGbrowse



# Generic Model Organism Database

GMOD - Chromium

GMOD  
Not secure | gmod.org/wiki/Main\_Page

Log in / create account



Page Discussion

Read

View source

View history

Search

Welcome to the **Generic Model Organism Database** project, a collection of open source software tools for managing, visualising, storing, and disseminating genetic and genomic data.

## Get Started

Read the [GMOD overview](#) for the big picture, or visit [GMOD Components](#) for a comprehensive list of GMOD tools. If GMOD looks promising for your needs, consider attending the next [GMOD community meeting](#).

## Get Support

GMOD support is available from several different sources. [Support](#) introduces each support option (this web site, [GMOD Mailing Lists](#), [Training and Outreach](#) activities (including [GMOD Schools](#)), and the [GMOD Help Desk](#)) and offers guidance on which one is the most appropriate for your question.

## Get Involved

As an open source project GMOD relies on the [donation of time and software](#) by groups and individuals. Contribution of new tools, adoption of existing ones, and [improving the documentation](#) are all welcome. [Existing](#) and potential users are encouraged to provide feedback via mailing lists or the [help desk](#). You can also attend project meetings.

## Popular GMOD Tools

See the [full list of GMOD components](#)



GMOD in the Cloud toolset



GBrowse: Genome annotation viewer



Galaxy: Data analysis & integration



Chado: Biological database schema



JBrowse: Super-fast genome annotation viewer



BioMart: Data mining system



WebApollo: browser-based annotation editor



MAKER: Genome annotation pipeline



GBrowse\_syn: Synteny viewer



Tripal: Chado web interface



InterMine: Data warehousing



CMap: Comparative map viewer



Pathway Tools: Metabolic, regulatory pathways



Canto: literature annotation tool

## GMOD is Social

Follow [The Tweet of GMOD](#)  
[@gmodproject](#)



Join other [GMOD users](#) on  
LinkedIn

Keep up with GMOD papers and  
contribute your own in the [GMOD  
group on Mendeley](#)



## GMOD News



[Prospecting for Proposals for GSoC 2017](#)  
[Computational Biologist GrainGenes](#)  
[Call for PAG Abstracts](#)  
[Now Hiring CTO Phoenix Bioinfo](#)  
[New GMOD Server](#)  
[GMOD-JBrowse 2016 Survey](#)  
[GCC2016](#)  
[2016 GMOD Meeting](#)  
[Prospecting for Proposals for GSoC 2016](#)  
[Prospecting for Proposals for GSoC 2015](#)

## Navigation

[GMOD Home](#)  
[Software](#)  
[Categories / Tags](#)  
[View all pages](#)

## Documentation

[Overview](#)  
[FAQs](#)  
[HOWTOs](#)  
[Glossary](#)

## Community

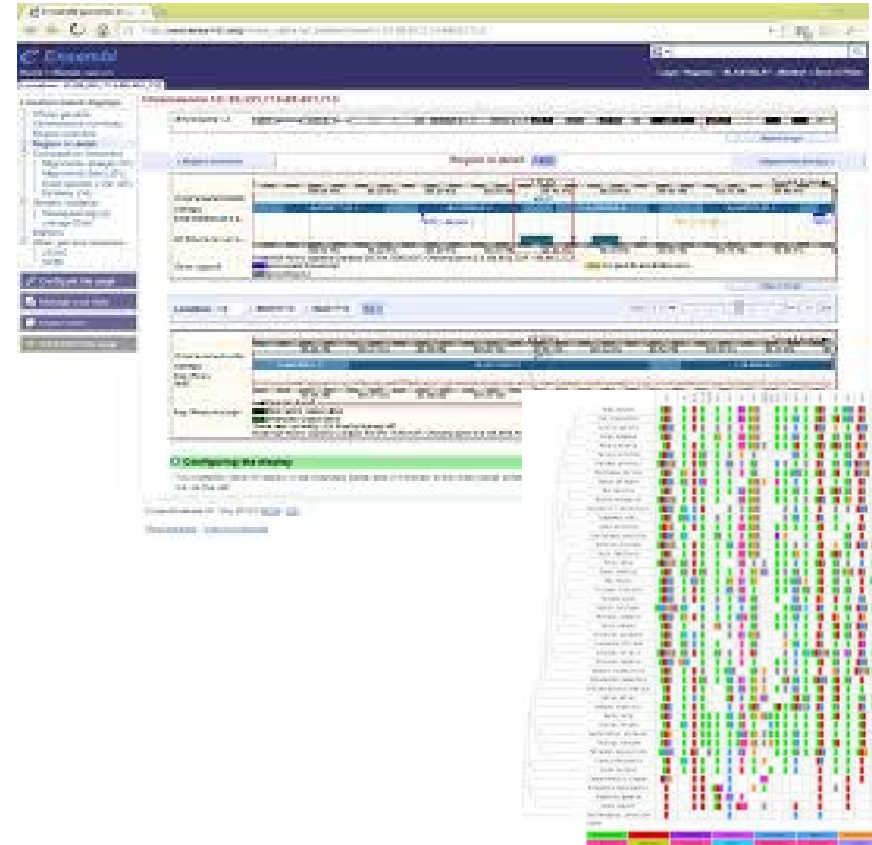
[GMOD News](#)  
[Training / Outreach](#)  
[Support](#)  
[GMOD Promotion](#)  
[Meetings](#)  
[Calendar](#)

## Tools

[What links here](#)  
[Related changes](#)  
[Special pages](#)  
[Printable version](#)  
[Permanent link](#)  
[Page information](#)  
[Browse properties](#)  
[Print as PDF](#)

# ENSEMBL

- European Nucleotide Service from EMBL
- Bazy w EBI niedaleko Cambridge
- Duża baza kodu w Perlu, schemat bazy danych w MySQL
- Licencja Apache
- Do niedawna niewiele genomów, obecnie ogromna liczba genomów i dużo informacji porównawczej





# Przeglądarki offline

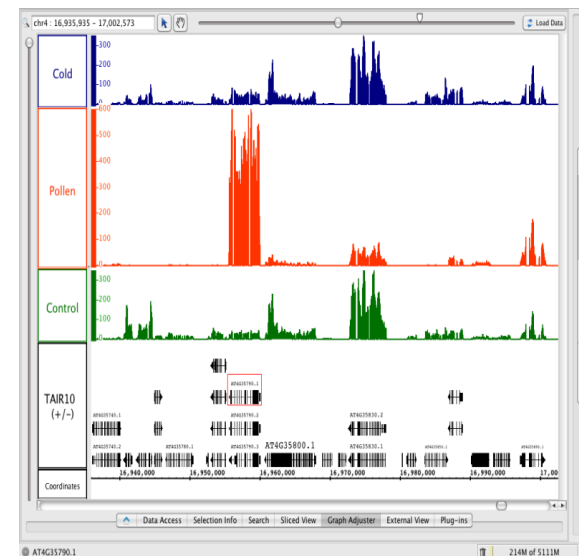
- Nie zawsze możemy używać przeglądarek online
- Np. możemy nie chcieć udostępniać danych na zewnątrz, albo możemy mieć za dużo danych, aby wysłać je na zewnętrzny serwer
- Rozwiązanie – przeglądarka korzystająca z danych lokalnych, wyświetlająca dane z dysku, ale pobierająca kontekst opisu genomu (geny, transkrypty, itp) ze zdalnego serwera (DAS)
- Typowo napisane w Javie – dostęp do grafiki i przenośność między Mac/Windows/Linux

# Distributed Annotation System

- BioDAS – projekt definiujący infrastrukturę (serwery i protokoły komunikacji z nimi) do rozpowszechniania anotacji genomowych
- Protokoły DAS 1 (tekst/XML), DAS 2 (dane binarne)
- Wspierany (w wersji 1) przez większość przeglądarek online i offline oraz “dostarczany” przez większość serwerów baz genomowych
- DAS 2 na razie nie jest istotną konkurencją dla alternatywnych rozwiązań (HTTP, track hubs)

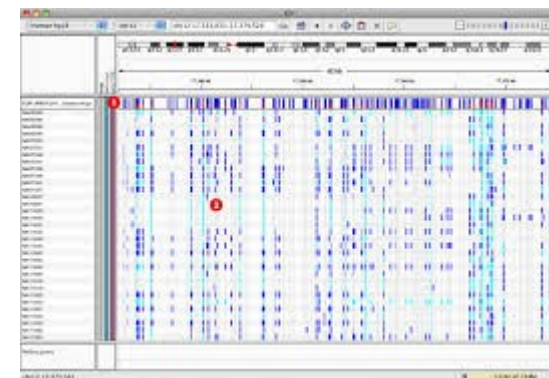
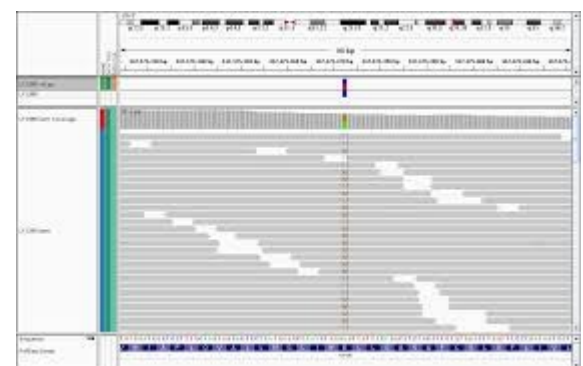
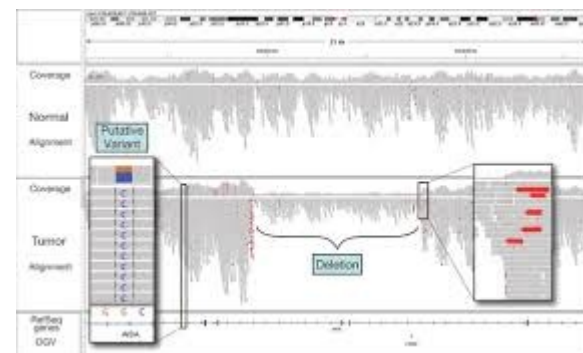
# Integrated Genome Browser

- Kod stworzony początkowo przez firmę Affymetrix do wizualizacji danych z macierzy “kafelkowych” (tiling arrays)
- “Porzucony” przez firmę i obecnie rozwijany w środowisku akademickim (UNC Charlotte)
- Academic free license
- Napisany w Javie, potem usunięto część kodu Affymetrixu



# Broad Integrative Genome Viewer

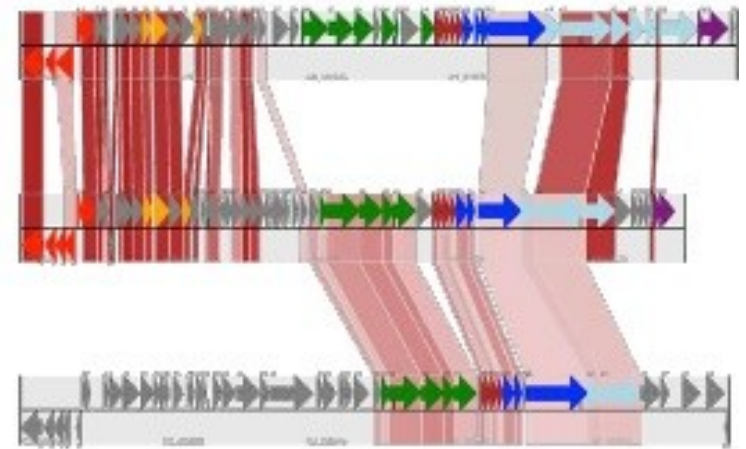
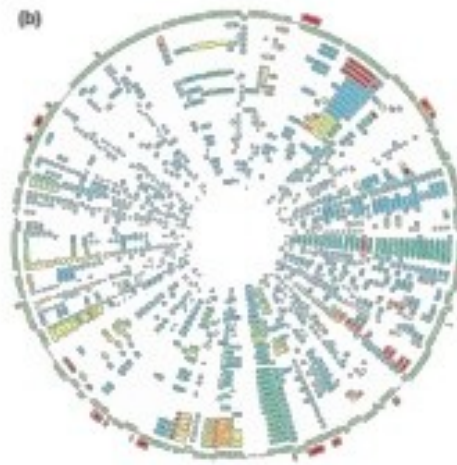
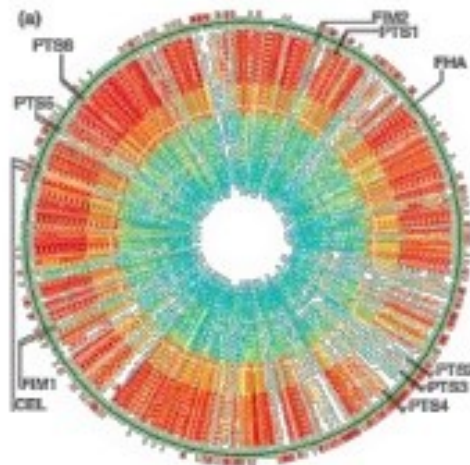
- W Zasadzie klon IGB, choć baza kodu zupełnie nowa
- Rozwój spowodowany brakiem możliwości komercjalizacji przez Broad (potencjalny konflikt z prawami Affymetrix) i problemami technicznymi
- Używany w częściowo komercyjnym genome space (obecnie też już chyba nie działającym)
- Licencja LGPL



# Genome Diagram



- Comparative genomics visualisation package
- Developed in 2003 for *Pba* sequencing, later incorporated into Biopython



<http://www.biopython.org>

Pritchard *et al.* (2006) *Bioinformatics* [doi:10.1093/bioinformatics/btk021](https://doi.org/10.1093/bioinformatics/btk021).